AFHRL-TR-90-92

(2)

AIR FORCE

HUMAN RESOURCES

APPLICATION OF GENERALIZABILITY THEORY TO THE AIR FORCE JOB PERFORMANCE MEASUREMENT PROJECT: A SUMMARY OF RESEARCH RESULTS

Kurt Kraiger

Department of Psychology
University of Colorado at Denver
1200 Larimer Street
Denver, Colorado 80204

Mark S. Teachout

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

July 1991

Interim Technical Report for Period October 1986 – December 1990

LABORATORY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

15

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

HENDRICK W. RUCK, Technical Advisor
Training Systems Division

HAROLD G. JENSEN, Colonel, USAF
Commander

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
July 1991

3. REPORT TYPE AND DATES COVERED
Interim Report – October 1986 – December 1990

4. TITLE AND SUBTITLE
Application of Generalizability Theory to the Air Force Job Performance Measurement Project: A Summary of Research Results

5. FUNDING NUMBERS
C – F41689-86-D-0052
PE – 63227F
PR – 2922
TA – 01
WU – 01

6. AUTHOR(S)
Kurt Kraiger
Mark S. Teachout

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Psychology
University of Colorado at Denver
1200 Larimer Street
Denver, Colorado 80204

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Training Systems Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFHRL-TR-90-92

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Generalizability theory was used to assess the psychometric quality of Walk-Through Performance Tests (WTPTs) and job proficiency ratings in eight occupational specialties. In addition, generalizability theory was used to determine whether proficiency ratings and job knowledge test scores were substitutable for the WTPTs. The results showed that both the WTPT scores and ratings within rating sources were generalizable (reliable), but that ratings were not generalizable over rating sources, and neither ratings nor job knowledge test scores ranked incumbents similarly to the WTPT.

14. SUBJECT TERMS
D-Study
G-Study
generalizability theory
job performance measurement
reliability

15. NUMBER OF PAGES
30

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

SUMMARY

The Air Force Job Performance Measurement (JPM) Project is a wide-scale effort to assess individual job proficiency. Incumbents are assessed via Walk-Through Performance Tests (WTPTs), job proficiency ratings, and (for some specialties), job knowledge tests.

This report summarizes three prior research efforts which supported the JPM Project by assessing the psychometric quality of both the WTPT and rating methods, and by examining the extent to which the ratings and the job knowledge tests are substitutable for the WTPT. In addition, the results of these analyses are compared across eight Air Force specialties to determine the extent to which judgments of measurement quality based on data collection to date are warranted. These issues are addressed primarily through the application of generalizability (G) theory. G theory is extensively reviewed, with reference to other applications in and out of the military. G theory is explained as a strategy for identifying whether scores assigned to individuals are dependable (or consistent) over conditions of measurement. For the rating data, the relevant conditions of measurement were rater sources, rating forms, and items or dimensions within particular forms. For the WTPT, relevant conditions of interest were assessment method (hands-on versus interview), tasks, and steps or items within tasks. For the substitutability issue, a third generalizability design was constructed with performance measures (WTPT scores, ratings, and job knowledge test scores) and tasks as the conditions of interest. Finally, for both the WTPT and rating measures, a subset of generalizability analyses known as D studies was employed to investigate the dependability of these measures under specific measurement conditions (e.g., a single rating source or a single WTPT method).

PREFACE

This work was performed under contract No. F41689-86-D-0052 with the Universal Energy Systems, Inc., as part of work unit number 77341301, Contributive Research in Performance and Training Evaluation. This work was initiated in response to Congressional and Air Force requests to validate selection and training systems against measures of job performance. This includes RPR 83-02, Improved Performance Measurement and Prediction; MPTN 89-11MP, Enlisted Selection and Classification; MPTN 89-13MP, Job Performance Measurement; and MPTN 89-15T, Development of Techniques for improved Training, Planning, and Evaluation.

The authors are grateful to Dr Jerry Hedge, Dr Hendrick Ruck, and Col Rodger Ballentine for their guidance and support throughout this effort.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

APPLICATION OF GENERALIZABILITY THEORY TO THE AIR FORCE JOB PERFORMANCE MEASUREMENT PROJECT: A SUMMARY OF RESEARCH RESULTS

I. INTRODUCTION

This report summarizes work performed between 1986 and 1989, applying generalizability (G) theory to data collected as part of the Air Force Job Performance Measurement (JPM) Project. The report is organized as follows: In Section I of this report, G theory is thoroughly introduced and discussed in the context of its value for evaluating the quality of job performance measures. Sections II and III describe the use of G theory to assess the reliability and substitutability of performance measures in eight Air Force occupational specialties. These two sections review work previously presented in Kraiger (1989, 1990a, 1990b) and Kraiger and Teachout (1990). The final section summarizes the implications of G theory results for the JPM Project and suggests new applications of G theory for training evaluation.

Introduction to Generalizability Theory

Generalizability theory was developed by Cronbach and his associates (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963) as an alternative to classical test theory. They replaced the notion of a single undifferentiated error term (basic to classical test theory) with a multifaceted error term which could be decomposed through carefully designed experiments. This is a considerably more realistic treatment of error variance because, in fact, measurement error can often be attributable to multiple, independent sources (e.g., occasion-to-occasion differences in subject attention, item difficulty, rater accuracy).

Within classical test theory, researchers have handled the problem of different sources of error by computing independent estimates of the effect of each source on test reliability. For example, the stability of test scores is examined by correlating two sets of scores obtained when a test is administered on two occasions, and the conspect reliability of a measure can be estimated by comparing the scores of two judges who evaluate the same responses. Though any *single* source of error variance can be estimated within classical test theory, relations among types of error variance cannot be estimated. For example, the direct effect of improving a test's internal consistency (by adding items) on the stability of test scores over time cannot be determined. Relationships among different kinds of measurement error are unclear and inestimable. Classical test theory leaves us with a fundamental paradox of a single true score but multiple estimates of true score variance depending on how error variance is defined. In other words, it allows us to estimate reliability, but cannot tell us what *the* reliability of a test is.

Multifaceted Approach of Generalizability Theory

In contrast to classical test theory, generalizability theory explicitly recognizes the existence of multiple sources of error variance and provides methods for simultaneously estimating each. In generalizability theory, the researcher identifies the factors ("facets" in G theory terminology) affecting measurement which are of the greatest interest or importance. Then, the researcher specifies a particular range of levels (or "conditions") of each factor for study. For example, if a researcher were studying the generalizability of performance ratings, facets might be the dimensions rated and the source of the ratings. Thus, the question is whether job incumbents' performance ratings vary as a function of the dimension rated or the rating perspective or style of the rater. Figure 1 is a Venn diagram illustrating potential sources of variance in a

study with raters and dimensions as facets. The σ_p^2 term provides the estimate of true score (or "universe" score) variance attributable to individual differences. All other components in the figure represent possible combinations of error variance.

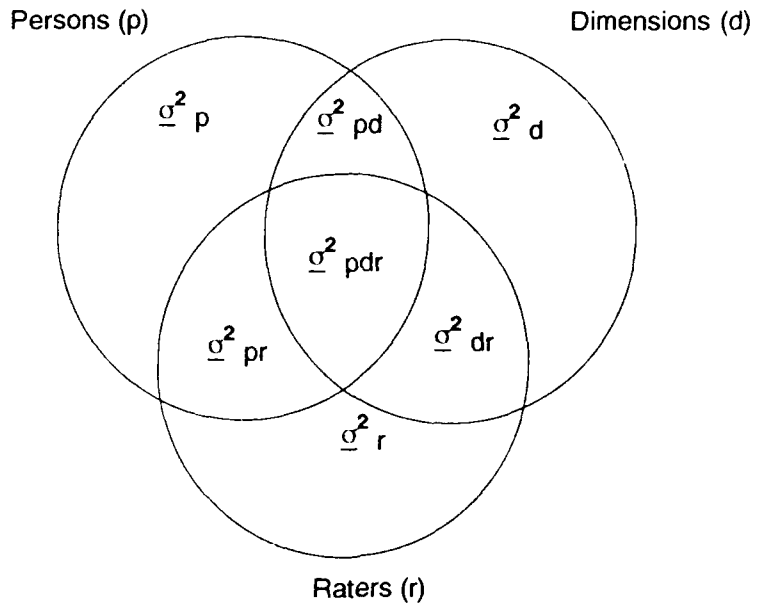


Figure 1. **Venn Diagram Illustrating a Two-Facet Fully Crossed Design for Analyses of Performance Ratings.**

A generalizability (G) study can be conducted to estimate the contribution of each facet to total score variance. In the present example, ratings for a large sample of incumbents would be obtained over random samples of dimensions and raters. Figure 2 provides a smaller example in which five incumbents were evaluated on four dimensions by two different raters. The tabled data in Figure 2 are fictitious, but illustrate several of the sources of error variance suggested by Figure 1. It should be noted that regardless of rater, each dimension yields a different average rating, with the highest mean generated by Dimension A and the lowest by Dimension D. In an actual G study, this pattern would result in a non-zero value for σ_d^2 and suggest that the ratings obtained by any incumbent would depend on the dimension assessed. Similarly, there is regularity in the ordering of the incumbents across dimensions and raters. Incumbent 3 is rated the highest by both raters, and both raters rated incumbent 2 as superior to incumbent 4. However, there are also differences in the rank-ordering of incumbents by raters. Compared to the first rater, the second rater reverses the relative positions of incumbents 1 and 2, and of incumbents 4 and 5. An actual G study of these data would result in a (desirable) large value for σ_p^2, indicating individual differences in performance when ratings are averaged over conditions of each facet. However, the analyses would also reveal an undesirable property of this rating system, a non-zero value for σ_{pr}^2, suggesting that each rater differentially ranks ratees. Similar assessments would be made for each of the other effects illustrated in Figure 1. For example, one might expect a non-zero value for σ_{pi}^2 in that both raters rated incumbent 2 higher than incumbent 1 on Dimension C but lower on Dimension D.

	Raters:									
	I					II				
	Dimensions					Dimensions				
Incumbent	A	B	C	D	$\bar{X}$	A	B	C	D	$\bar{X}$
1	4	4	2	4	3.5	3	2	1	3	2.3
2	4	2	4	2	3.0	4	3	4	2	3.3
3	5	5	4	4	4.5	4	4	4	4	4.0
4	3	1	2	2	2.0	4	2	3	2	2.8
5	4	3	4	2	3.3	1	2	3	1	1.8
$\bar{X}$ =	4.0	3.0	3.2	2.8	3.5	3.2	2.6	3.0	2.4	2.8

Figure 2. **Sample Data for a Two-Facet Fully Crossed Design for Analyzing Performance Ratings.**

The multifaceted treatment of error variance by generalizability theory should be clearer at this point. For any set of persons, the total observed score variance is represented by the left full circle in Figure 1. Generalizability theory partitions that variance into individual difference variance (σ_p^2) and multiple, distinct sources of error variance (in this example, σ_{pr}^2, σ_{pd}^2, and σ_{prd}^2. Just as classical test theory provides a reliability coefficient conceptualized as the ratio of true score to total variance (σ_t^2/σ_x^2), generalizability theory provides a generalizability coefficient, εP^2, equivalent to σ_p^2/σ_x^2 or $\sigma_p^2/(\sigma_p^2 + \sigma_{pr}^2 + \sigma_{pd}^2 + \sigma_{prd}^2)$.

Variance components computed from a G study (e.g., σ_p^2) represent estimated variance about universe scores for average single observations (e.g., an average person evaluated on an average task by an average administrator). These variance components enable a comparison of the relative contributions of error variance sources or the computation of summary generalizability coefficients. A third use of G study variance components is as input to decision (D) studies. D studies better reflect how an organization uses a measurement instrument.

Decision Studies

Although a G study establishes the relative effects of different sources of variance, it does so under conditions that may not reflect the intended use of the measurement instrument. Organizational decision-makers may wish to forecast how well a measure may perform under conditions which vary from the current context. A D study can be conducted to assess the specific characteristics of a measurement instrument in a particular decision-making context. For example, a D study could be conducted to determine the reliability of the Walk-Through Performance Test (WTPT) if more test administrators were added, or fewer tasks were sampled. In any D study, two critical specifications are (a) the universe of generalization, and (b) whether conditions within facets are to be treated as fixed or random (Gillmore, 1979, 1983).

The concepts of the universe of admissible observations and the universe of generalization are closely related. The universe of admissible observations refers to the facets that one decides to include in a G study and to the range of conditions which can be sampled from each. The universe of generalization in a corresponding D study may be no larger than the universe of admissible observations. That is, it cannot contain facets missing from the universe of admissible observations, nor can it contain a broader range of conditions. The universe of generalization may be smaller. For example, if the universe of admissible observations includes all tasks performed by Air Traffic Control Operators, a smaller subset of tasks such as all

monitoring tasks or all tasks accomplishing a single objective may be specified at the D study level.

The second consideration is whether each facet is to be treated as fixed or random. A fixed facet means that the range of conditions sampled at the G study level exhaust the range of conditions of interest to the researcher. A random facet means that the conditions of the facet represent a random sample from a larger set of admissible conditions. In practice, one must at least be willing to assume that the conditions sampled could be replaced with other elements of some larger set of possible observations without affecting the universe score (Shavelson & Webb, 1981). For example, the tasks assessed on the WTPT or a task-level rating form may be considered a random facet if there were other tasks which could have been included on these measures. When a random facet is specified, generalization is not limited to the set of D study conditions but instead, extends to the entire range of admissible observations.

When a facet is fixed, generalization is limited to the range of conditions included in the D study. Operationally, fixed effects may be treated in one of two ways. First, separate variance components for the other facets may be computed within each level of the fixed facet. For example, if WTPT methods (hands-on or interview) were a fixed facet, separate variance components involving other measurement facets could be computed for both the hands-on and interview scores. This method is recommended when the G study variance for the fixed facet is relatively large. A second strategy is to calculate a single summary score over all conditions of the fixed facet and apply this summary score to subsequent questions of generalizability. Resulting generalizability coefficients will be larger because possible variance due to the interaction of persons and the facet has been removed by averaging scores over the facet. However, this improvement would hold only for the particular fixed set of conditions.

It should be noted that considerations of fixed and random facets occur at the D study level, not at the G study level. For computing G study estimates of variance components, all facets are treated as random (i.e., all are estimated). In subsequent D studies, the variance components are set to zero if the generalizability of average scores for that facet is of interest. Shavelson (1986) has recommended inspection of the G study variance components for the fixed facet as a means of deciding how to treat the facet for D study analyses. If the variance component for the fixed facet is large, it may be inappropriate to average over its conditions. In such cases, Shavelson recommends separate D study analyses for other facets within each level of the fixed facet.

In addition, the number of D study conditions for each facet are not restricted to the number sampled in the G study. Rather, the investigator can systematically vary the number of conditions to forecast changes in generalizability. G study variance component estimates are actually average effects for single occurrences of each facet. That is, a G study variance component represents measurement error when only a single level of the facet is used. Because measurement error decreases as scores for the object of interest are averaged over multiple levels of a facet, D study estimates of variance components are computed to estimate the actual degree of error variance under conditions of multiple operationalizations of a facet. D study estimates are computed by dividing the G study variance component estimates by the number of conditions specified in the D study. Computing different D study estimates for differing numbers of conditions allows the researcher to predict how dependable a measure would be under a variety of measurement conditions.

Finally, it should also be noted that Cronbach et al. (1972) specified two different error terms, depending on the purpose of measurement. When measures are used for relative decisions, scores are used only to rank-order persons (e.g., as a criterion in a test validation study). In this case, errors in persons' actual scores do not matter, as long as these errors are equivalent for each person in the sample. Problems occur only when errors exist for

some persons but not others. Thus, if a test consisted of a particularly difficult sample of items, variance due to items would not be considered error because test difficulty alone would lower all scores but preserve the rank order. However, the person-by-item variance component would be considered error in that this interaction means that some persons are affected more than others by item difficulty. Generally, the error term for relative decisions includes all variance components which represent an interaction of a facet with persons.

In contrast, absolute decisions are decisions made in reference to a person's actual score. An example would be a college admissions policy which accepted all students who achieved a particular score on their board examinations. For absolute decisions, all sources which affect the level of the score are included in the error term. Thus, variance due to items would be included because a particularly easy or difficult sample of items would affect a person's likelihood of passing. In general, the error term for absolute decisions includes all variance components other than that component which constitutes the universe score variance (typically persons). Typically, specifications of relative or absolute error terms are made at the D study level.

To summarize, a single large-scale G study is conducted to estimate variance components for each effect in a model. From these estimated variance components, the researcher can generate numerous sets of D study variance component estimates and generalizability coefficients, depending upon how the measuring instrument is to be used. These D study results are of greater interest to decision-makers because they reflect realistic or intended measurement conditions.

Applications of Generalizability Theory

To further explain generalizability theory, several previous applications from the field of industrial/organizational (I/O) psychology will be discussed. One is a study by Webb, Shavelson, Shea, and Morello (1981) of the generalizability of General Educational Development (GED) ratings by experienced job analysts. Such ratings are often used to estimate training requirements or refer persons to job training programs. The study employed a three-facet design, with raters nested within offices and crossed with jobs and occasions. Seventy-one raters were nested within one of 11 different field centers; 27 jobs were rated on two different occasions. Jobs were the object of measurement and were not considered a facet. Separate analyses were performed for each of three GED rating scales--reasoning development, mathematics development, and language development.

Favorable levels of generalizability were reported for the GED ratings. Inspection of the variance components showed that differences in jobs (universe score variance) accounted for the largest variance in ratings. Thus, considering ratings from an average rater at an average center on one occasion, generalizability coefficients ($\underline{\varepsilon}P^2$) ranged from .53 to .67 over the three scales. D study data showed that the generalizability coefficients ranged from .79 to .85 for the mean rating of four raters. The largest sources of undesirable error variance were jobs crossed with raters within centers, and the residual error term. The former component indicated idiosyncratic perceptions of certain jobs by raters at certain centers. That is, errors of this nature would result if raters at one center perceived the GED requirements of one job differently than did raters at other centers.

Applications of generalizability theory to performance measurement studies are rare. In one earlier study, Littlefield, Murrey, and Garman (1977) examined the generalizability of faculty ratings of third- and fourth-year dental students. Ratings were made on five general dimensions of noncognitive skills. The ratings were collected from 31 faculty members on 12 students during one phase and from 16 faculty on 5 students during another phase. Each phase was considered a separate G study as each reflected a different set of measurement conditions

(differing numbers of raters). Students were the object of measurement and the facets of generalization were raters (faculty) and rating scales. Separate D study analyses were also conducted, with the scales treated as fixed or random (i.e., a mean scale score was used). The generalizability of students' ratings across raters and/or scales was quite high. The generalizability coefficients were .92 and .83 for the two phases of ratings. When scales were considered fixed and the generalizability of the mean score across raters was computed, the generalizability coefficients increased to .95 and .86. Thus, about 90% of the variance in scores could be attributed to universe score variance, or individual differences. Simulated D study results were also computed for ratings obtained from one rater versus two raters. As expected, generalizability coefficients were considerably lower, ranging from .53 to .61 for one rater and from .68 to .76 for two. Littlefield et al. concluded that at least two raters were necessary for dependable ratings.

Two other studies derive from military settings. In one, McHenry, Hoffman, and White (1987) conducted a generalizability analysis of performance ratings for 7,045 soldiers in 19 Army jobs. For their analyses, rater type (peers and supervisors) and rating scale were the facets of interest. Analyses were performed for each job and within each of three previously identified general performance factors (effort/leadership, personal discipline, and physical fitness/military bearing). When scores were averaged over rater type, generalizability coefficients were very high for two of the three performance factors. Inspection of the individual variance components revealed that generalizability was lower on the third factor (physical fitness/military bearing) because of a large interaction between rating scales and ratees. In other words, ratees were differentially ranked across scales comprising that factor.

Finally, a recent article by Webb, Shavelson, Kim, and Chen (1989) summarized the results of separate generalizability analyses on four different measures of performance for the job of machinist mate--hands-on performance tests, paper-and-pencil job knowledge tests, task-level performance ratings, and global performance ratings. Because of their similarity to analyses presented later in the present report, two of the Webb et al. investigations will be explained more fully. For their study of the generalizability of hands-on performance tests, Webb et al. examined two crossed facets--tasks and observers. The observers were two trained examiners, who scored the machinist mates on all tasks. The tasks were 11 duties commonly performed in the engine room. Because of thorough training, the observers produced identical rank-orderings of mates on all tasks (i.e., $\underline{\sigma}_{po}^2 = .00$). Acceptable levels of generalizability ($\underline{\varepsilon P}^2 > .70$) were achieved by averaging scores over at least 11 tasks.

The investigation of performance ratings used tasks and rater types (self, peer, and supervisor) as rating facets. Large values were found for both $\underline{\sigma}_{pt}^2$ (indicating that machinist mates were differentially ranked by job tasks) and $\underline{\sigma}_{pr}^2$ (indicating that mates were differentially ranked by rater types). Acceptable levels of generalizability were obtained only by assessing at least 11 tasks and by assuming that rater types were fixed (thus averaging ratings over all three sources).

Applicability to the JPM Project

The application of generalizability theory to the Air Force JPM project was recommended on several grounds (Kraiger, 1989). First, G theory offers a more versatile and realistic portrayal of measurement error than does classical test theory. For any measurement system, multiple sources of error interact to threaten the fidelity of measurement. Only G theory enables simultaneous estimation of each potential threat and of the interactive contributions of multiple sources of error variance. Second, generalizability theory forces the researcher to explicitly address measurement issues that are often otherwise ignored. Examples of these issues include the precise conditions of measurement which may affect scores and whether an instrument is to be used for relative or absolute decision-making. Third, D study analyses

permit decision-makers to predict reliability of measurement under a host of possible measurement conditions not currently employed. Finally, generalizability theory was recommended by the National Academy of Sciences committee monitoring the Joint-Service Job Performance Measurement Project (Wigdor & Green, 1986).

G theory was applied to assessments of the psychometric quality of performance ratings in four Air Force specialties (AFSs) in Kraiger (1990a), and an additional four AFSs in Kraiger (1990b). This research is summarized below. First, the types of variables and the questions addressed by generalizability theory will be explained in greater detail.

G Theory Analyses in Eight AFSs

Within the JPM Project, incumbent work proficiency is assessed using three methods of measurement: WTPTs, job proficiency ratings, and job knowledge tests. The WTPT is a work sample test composed of hands-on performance tests and interviews. The hands-on format requires airmen to perform a series of actual tasks under the careful observation of a highly trained test administrator. With the interview, incumbents describe in detail the steps they would perform to accomplish similar tasks. Proficiency ratings consist of performance evaluations collected on each of four different rating forms by three different sources: incumbents, one to three peers, and an immediate supervisor. Both the WTPT and the rating forms are described in greater detail in Hedge and Teachout (1986). Incumbents in four specialties were also administered job knowledge tests. These tests require incumbents to answer multiple-choice questions regarding critical on-the-job tasks. Additional details on job knowledge tests are provided in Bentley, Ringenbach, and Augustin (1989).

Five generalizability designs were employed to investigate the psychometric quality of the performance ratings. Each is explained in the subsequent section.

Generalizability Designs

Generalizability of Rating Data. Two designs were used to investigate the generalizability of performance ratings over measurement conditions. In one, facets of interest were rating sources (incumbents, peers, and supervisors), forms (task-level, dimensional, global, and Air Force-wide), and the number of items (or scales/dimensions) nested within forms. Together, these facets generated 11 potential sources of variance in scores.

The first facet was rating sources; incumbents, peers, and supervisors were the conditions of the facet. The sources can be considered random samples of a larger universe of possible sources which could be used to assess ratee performance. When airmen were rated by more than one peer, only a single randomly selected rating was used in order to balance the design. The second facet was rating forms, with task-level, dimensional, global, and Air Force-wide forms as the conditions of the facet. These can be considered random samples of a larger universe of possible forms which could be used to assess ratee performance. The final facet was the individual items which comprised each form. Again, the items on any one form can be considered a random sample of possible items which could constitute that form. Items were nested within forms because individual items or scales vary from form to form.

Because the forms differed as to their level of detail, the number of items comprising a form varied considerably. To balance the number of items over forms (and avoid unbiased mean square estimates; see Searle, 1971), analyses were conducted with two randomly selected items from all four forms and with x number of items from all forms except the two-item global form, where x was the number of items on the dimensional form (the next smallest form).

Results from both analyses were similar, and yielded comparable conclusions regarding the generalizability of ratings. Only the results of the three-form analyses are presented in this report because these contain less sampling error than the four-form analyses.

Initial analyses yielded extremely large variance components in all specialties for the interaction of sources of ratees (σ_{ps}^2), indicating that airmen were differentially ranked by different sources. This finding, coupled with theories in I/O psychology which treat such differences as valid (Borman, 1974; Klimoski & London, 1974; Kraiger & Teachout, 1990), led to the specification of a second generalizability design. Here, separate analyses were conducted for each rating source. Facets were forms and items nested within forms.

Generalizability of WTPT Data. Another set of generalizability analyses focused on the WTPT. These analyses were accomplished by two designs, each with three facets. The first facet was the assessment method, with the hands-on and interview methods as the conditions of the facet. The second facet was the tasks that were measured by both the hands-on and the interview methods. Typically, a WTPT consisted of 20 to 25 tasks. For each specialty, these tasks can be considered random samples of a larger possible universe of tasks which could comprise the WTPT. For purposes of analysis, there were two possible generalizability designs investigating variance due to tasks. For each specialty, there were three types of tasks included in the WTPT: tasks common to both the hands-on and the interview components, tasks unique to the hands-on component, and tasks unique to the interview component. Thus, common tasks were assessed by both methods, whereas unique tasks were assessed by one WTPT method but not the other. One analysis (the "crossed design") included only the common tasks and treated tasks as crossed with methods, because each task is assessed by each method and each method includes all tasks. A second analysis (the "nested design") included the unique tasks and treated tasks as nested within methods, because tasks differed across methods of the WTPT. To maximize the number of tasks analyzed (thereby reducing sampling error in the entire design), analyses were conducted with both common and unique tasks nested within methods. For example, eight unique tasks and six common tasks may have been analyzed as nested within a method even though these common tasks were not really nested. These analyses produced similar results.

The final facet of interest was the number of items or steps comprising individual tasks on the WTPT. In scoring the WTPT, a person's score on a task is determined by the number of correct steps completed on the task. Items were treated as nested within tasks because they were in fact different for each task on the WTPT. For each task, the items can be considered random samples of larger possible universes of possible items. Again, the items facet for the WTPT was unbalanced in that the number of steps for a task ranged from as few as 4 to over 30. To balance the items facet, tasks with only a few items were dropped from the analyses, and items were randomly selected from longer tasks to match the number of items in the smallest of the remaining tasks. For example, for the Information Systems Radio Operator, tasks with less than six items were not analyzed, and six items were randomly sampled for all tasks with more than six steps. For two specialties, AFS 122X0 and AFS 324X0, only unique tasks remained after the elimination of tasks with a small number of steps. Consequently, only the nested design was used for these two specialties.

To summarize, in the WTPT crossed design, methods and tasks were crossed and steps were nested within tasks. Eleven sources of variance could be estimated from this design. In the nested design, steps were nested within tasks, which were nested within methods. Seven sources of variance could be estimated from this design.

Substitutability Design. The fifth generalizability design assessed the extent to which proficiency scores generalized over the three primary measurement methods: ratings, WTPT, and job knowledge tests. All analyses were conducted with a fully crossed design; the facets of interest were evaluation methods and tasks. In four AFSs (AFS 426X2, AFS 272X0, AFS

328X0, and AFS 492X1), the method facet consisted of two conditions--task-level ratings and overall WTPT scores. In four additional specialties for which complete data were available (AFS 122X0, AFS 732X0, AFS 324X0, and AFS 423X5), the method facet consisted of all three evaluation methods. Analyses were conducted separately for ratings provided by each of the three rating sources. Only results for supervisory ratings are presented in this report, as these ratings resulted in the highest level of generalizability.

The number of tasks analyzed was equal to the smaller number of tasks on either the hands-on or interview component of the WTPT for a specialty. An equivalent number of tasks were randomly sampled from the other WTPT component, from the task-level rating form, and (when available) from the job knowledge test. For the job knowledge data, several questions tapped the knowledge(s) required for each task. Task scores were computed by determining the percentage of questions correct within each task sampled.

D Study Analyses

D study analyses were conducted with G study estimated variance components as input. D study results included variance components for individual effects, total universe score variance (variance due to individual differences), relative error variance ($\underline{\sigma}^2_{\delta}$, equal to the sum of all effects which contain p and at least one other index), absolute error variance ($\underline{\sigma}^2$, equal to the sum of all effects in the design except $\underline{\sigma}^2_{p}$), and their associated generalizability coefficients ($\underline{\varepsilon P}^2$, for relative decisions; θ, for absolute decisions).

Conditions in the D study were specified as follows. All facets were treated as random in the ratings and substitutability design. For analyses of the WTPT, the methods facet was analyzed as both random and fixed. A fixed facet implies that the conditions observed in the G study exhaust the range of possible conditions of interest to the organization and that the organization intends to use an average or total score over conditions of the facet.

Secondly, the number of conditions observed for each facet were systematically varied at the D study level to estimate generalizability under measurement conditions of various levels of complexity. For example, generalizability coefficients were computed for the multiple combinations of possible sizes of the WTPT (e.g., 10 items on 10 tasks with one WTPT method, or 15 items on 5 tasks with two methods). Operationally, a D study variance component is adjusted by dividing the variance component by the number of conditions of any facet indicated by its subscript. For example, the G study estimate for $\underline{\sigma}^2_{i:f}$ would be divided by 24 if eight items on each of three forms were specified.

To distinguish D study estimates from unitary G study values, D study facets were noted by capital letters in the subscript. However, the "p" associated with individuals remains lowercase because persons are not treated as a facet in these analyses. Thus, the G study effect $\underline{\sigma}^2_{i:f}$ is indicated as $\underline{\sigma}^2_{I:F}$ at the D study level, and $\underline{\sigma}^2_{pm}$ is indicated as $\underline{\sigma}^2_{pM}$ at the D study level (Brennan, 1983; Brennan & Kane, 1979).

II. METHOD

Sample

Measures of work proficiency were collected from first-term airmen in eight different specialties: Jet Engine Mechanic (AFS 426X2), $\underline{n}$ = 255; Air Traffic Control Operator (AFS 272X0), $\underline{n}$ = 172; Avionic Communications Specialist (AFS 328X0), $\underline{n}$ = 98; Information Systems Radio Operator (AFS 492X1), $\underline{n}$ = 158; Aircrew Life Support (AFS 122X0), $\underline{n}$ = 216; Personnel Specialist (AFS

732X0), n = 218; Precision Measuring Equipment Specialist (AFS 324X0), n = 138; and Aerospace Ground Equipment (AFS 423X5), n = 264. For all specialties, incumbent performance was measured via the WTPT and proficiency ratings. Job knowledge tests were also administered to incumbents in the latter four specialties. The WTPT was administered on the job site and required incumbents to perform and/or describe the sampled tasks. Performance was observed by a carefully trained administrator, who recorded whether or not incumbents executed (or described) the correct steps to accomplish the task. In addition, incumbents were rated, on each of four rating forms, by themselves, by one or more peers, and by their immediate supervisor. The job knowledge test consisted of multiple-choice questions designed to assess an understanding of the tasks completed on the WTPT or rated on the forms. Data generated by these measures as part of the JPM project were analyzed using GENOVA, a Fortran-based computer program especially designed for generalizability analyses (Crick & Brennan, 1982). Specifications were provided to the program to represent each of the G study and D study designs discussed above.

III. RESULTS

Ratings Design

G Study Results. Estimated G study variance components for each effect are presented in Table 1 for eight occupational specialties. The sizes of the relative variance components were similar across specialties. In each specialty, the first and second largest variance components were the residual term ($\sigma^2_{ps\,(i:f)}$) and the interaction of ratees and sources (σ^2_{ps}). The σ^2_p term, universe score variance, was the third largest term for all specialties except Information Systems Radio Operators and Personnel Specialists. This term varied from .047 to .151. Similar narrow ranges across specialties can be seen for most other terms. Only a few terms showed considerable variation across specialties. The main effect for rater sources, (σ^2_s) was near zero in six specialties, but substantially larger for Air Traffic Control Operators and Personnel Specialists. This effect was largely due to low mean supervisory ratings for Air Traffic Control Operators and exceptionally high self-ratings by the Personnel Specialists.

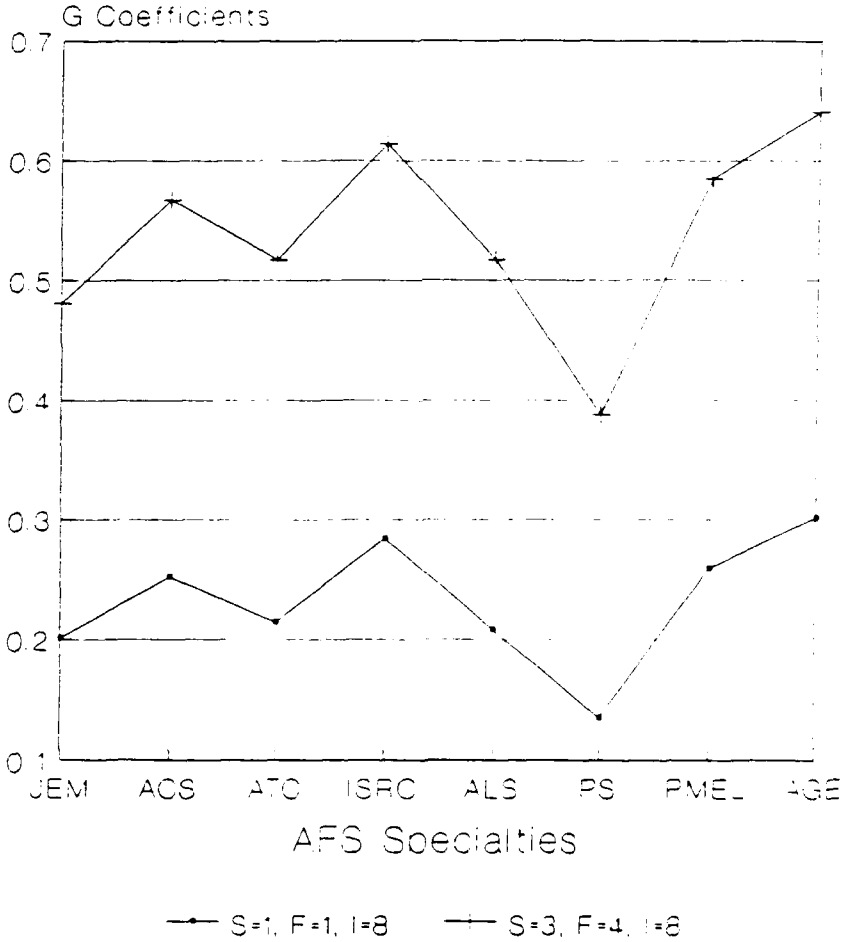
Table 1. **Estimated Variance Components for G Study of Rating Variables with Three Forms**

	Job							
Effect	JEM σ^2	ACS σ^2	ATC σ^2	ISRO σ^2	ALS σ^2	PS σ^2	PMEL σ^2	AGE σ^2
p	.151	.120	.118	.133	.088	.047	.087	.122
s	.015	.015	.036	.001	.010	.041	.010	.016
f	.001	-.001	-.017	-.009	.001	.002	-.005	-.006
i:f	.015	.031	.040	.025	.039	.045	.049	.054
ps	.186	.173	.208	.173	.193	.172	.140	.160
pf	-.003	-.030	-.009	.021	.028	.023	.027	.022
sf	.001	-.008	.000	.003	.000	.000	.001	.000
psf	.016	-.018	.010	.036	.061	.043	.033	.048
p(i:f)	.057	.106	.066	.089	.074	.094	.065	.055
s(i:f)	.004	.019	.000	.002	.005	.005	.002	.007
ps(i:f)	.293	.330	.285	.306	.353	.395	.322	.359

Note. JEM = Jet Engine Mechanic, ACS = Avionic Communications Specialist, ATC = Air Traffic Control Operator, ISRO = Information Systems Radio Operator, ALS = Aircrew Life Support, PS = Personnel Specialist, PMEL = Precision Measuring Equipment Specialist, AGE = Aerospace Ground Equipment, p = persons, s = sources, f = forms, i:f = items within forms.

The σ^2_{psf} term was considerably lower for three specialties (AFS 426X2, AFS 272X0, and AFS 328X0) than in the other five. This pattern suggests that in these three specialties, ratees were ranked similarly regardless of which combination of rater source and form was used, but in the other five specialties, the interaction of form and source affected a ratee's relative ranking. For example, an incumbent in Aerospace Ground Equipment might be ranked above a co-worker by peers using one form, but ranked below that co-worker by a supervisor using a different form.

D Study Results. Summary generalizability (G) coefficients for relative decisions ($\underline{\varepsilon P^2}$) for all specialties are presented in Figure 3 for two sets of measurement conditions: a single source using a single 8-item form (representing typical organizational operationalizations of rating methods), and three sources using four 8-item forms (the D study which best approximates the actual measurement conditions on the JPM). The generalizability coefficient represents the proportion of observed score variance which is attributable to universe score variance or individual differences. It indicates the overall dependability of measures under a particular set of conditions.



Note. JEM = Jet Engine Mechanic, ACS = Avionic Communications Specialist, ATC = Air Traffic Control Operator, ISRO = Information Systems Radio Operator, ALS = Aircrew Life Support, PS = Personnel Specialist, PMEL = Precision Measuring Equipment Specialist, AGE = Aerospace Ground Equipment, S = sources, F = forms, I = items.

Figure 3. **G Coefficients for Performance Rating Data for Eight Occupational Specialties.**

As shown in Figure 3, rating measures were more reliable when ratings were averaged over multiple sources and multiple forms. With a single source using a single 8-item form, G coefficients ranged between .135 and .302. In contrast, by averaging scores over all three sources and four forms, the generalizability scores ranged from .388 to .641, with most values above .500. Notably, the generalizability coefficients are comparable across the specialties, except that the values for Personnel Specialists were considerably lower than those in the other seven specialties.

According to the design, generalizability coefficients may be increased by increasing the numbers of rating dimensions, forms, and (particularly) sources. For example, the $\underline{\sigma}^2_{p\ (1:f)}$ term is small, but non-trivial in the G study results presented above. By averaging ratee scores over multiple items and/or forms, this undesirable source of variance can be virtually eliminated at the D study level. Similarly, averaging over multiple sources reduces the $\underline{\sigma}^2_{ps}$ and $\underline{\sigma}^2_{psf}$ terms. However, the ratee-by-source interaction term is still large, even when ratings are averaged over three sources. This source of variance is the greatest threat to the generalizability of the performance ratings.

Within-Source Analyses

Secondary analyses were performed within each rater source for each specialty. G study results and the D study G coefficients (for a single 8-item form) are displayed in Table 2. The generalizability coefficient is also displayed in Figure 4 for each source.

Variance components and G coefficients were again consistent over specialties. At the D study level, variance due to individual differences ($\underline{\sigma}^2_p$) was substantial for each source within each specialty, whereas most other sources of variance were negligible.

Fairly large D study generalizability coefficients were obtained, even with a single 8-item form. The majority of generalizability coefficients under these conditions ranged from .660 to .750 across sources and specialties. The largest G coefficients were most often obtained for supervisory ratings, though larger coefficients were found in two other specialties for peer ratings (Avionic Communications Specialists) and self-ratings (Air Traffic Control Operators).

G Study Results, WTPT Data

Results of the G study analyses across specialties are presented in Table 3 (for the crossed design) and Table 4 (for the nested design). There is considerably more variability across specialties in the analyses of WTPT scores than of proficiency ratings. For example, variance due to individual differences ($\underline{\sigma}^2_p$) ranged from .006 for Avionic Communications Specialists to .032 for Information Systems Radio Operators. Likewise, the residual term ($\underline{\sigma}^2_{pm\ (i:t)}$) was considerably larger in the Jet Engine Mechanic and Aerospace Ground Equipment specialties than in the others. The $\underline{\sigma}^2_{pm}$ and $\underline{\sigma}^2_{pmt}$ terms were relatively small and consistent across specialties, but considerable variation in estimates was found for the $\underline{\sigma}^2_{pt}$ and $\underline{\sigma}^2_{p\ (i:t)}$ terms. The estimate for the person-by-task interaction was substantially larger for Avionic Communications Specialists, Air Traffic Control Operators, and Information Systems Radio Operators than in the other specialties. This indicates that incumbents in these three specialties were differentially ranked on performance, depending on the task. Variability was also found for the interactions of persons and items nested within tasks. This term was near zero for all specialties except Avionic Communications Specialists and Air Traffic Control Operators.

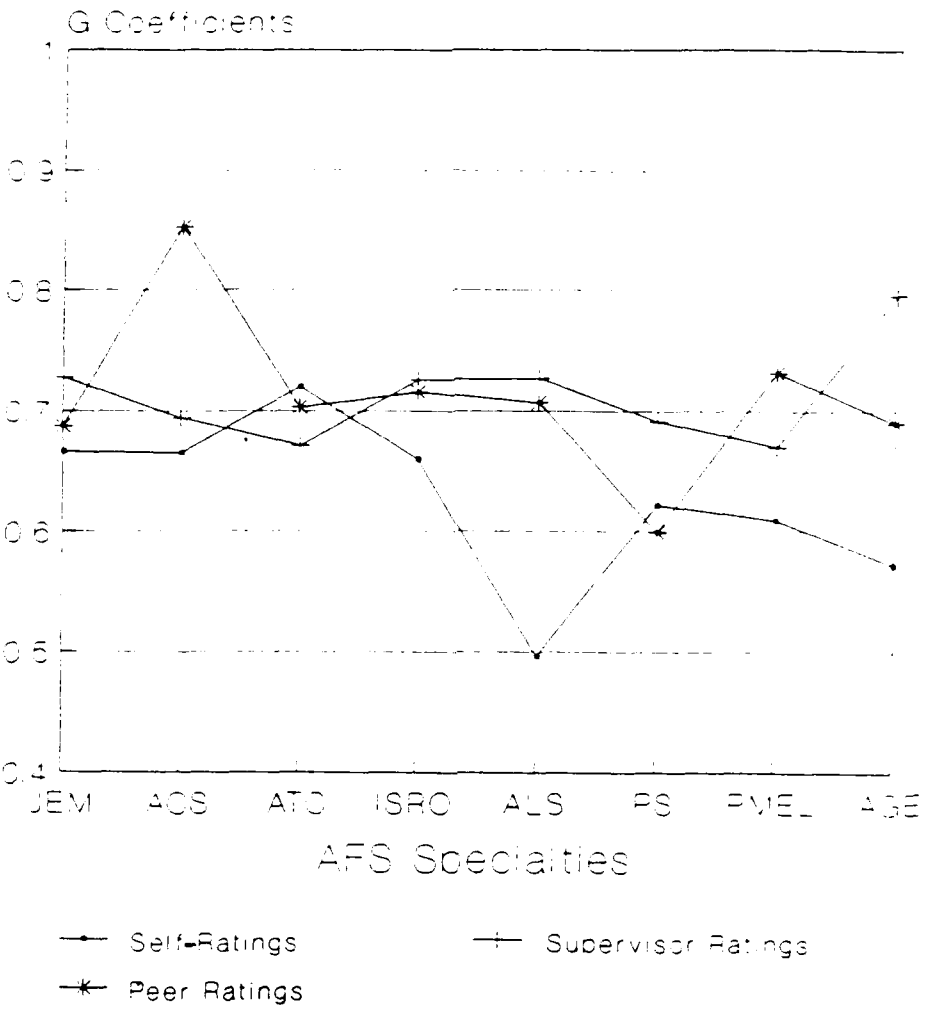
Table 2. **G and D Study Results for Within Source Analyses**

Source: Effect	Job JEM σ^2	ACS σ^2	ATC σ^2	ISRO σ^2	ALS σ^2	PS σ^2	PMEL σ^2	AGE σ^2
Self:								
p	.192	.161	.218	.219	.145	.149	.147	.163
f	.035	.014	.011	.030	-.006	.001	.007	.008
i:f	.025	.034	.021	.019	.066	.034	.062	.087
pf	.048	.030	.038	.073	.094	.034	.044	.065
p(i:f)	.351	.415	.376	.314	.473	.451	.403	.464
εP^2 when f = 1, i:f = 8	.666	.665	.720	.660	.496	.622	.609	.572
Supervisor:								
p	.375	.275	.312	.289	.363	.271	.279	.400
f	.026	.038	.014	.062	.002	-.006	-.007	-.005
i:f	.026	.026	.029	.035	.031	.068	.050	.049
pf	.097	.069	.103	.063	.087	.070	.090	.061
p(i:f)	.346	.420	.400	.373	.396	.456	.382	.383
εP^2 when f = 1, i:f = 8	.728	.694	.671	.726	.727	.691	.670	.796
Peer:								
p	.265	.357	.291	.282	.337	.234	.256	.282
f	.047	.051	.019	.056	.006	.010	-.004	-.005
i:f	.024	.017	.031	.015	.035	.047	.042	.046
pf	.077	.020	.075	.072	.088	.093	.047	.083
(i:f)	.350	.328	.387	.314	.411	.562	.374	.396
εP^2 when f = 1, i:f = 8	.687	.853	.703	.716	.707	.599	.732	.690

Note. JEM = Jet Engine Mechanic, ACS = Avionic Communications Specialist, ATC = Air Traffic Control Operator, ISRO = Information Systems Radio Operator, ALS = Aircrew Life Support, PS = Personnel Specialist, PMEL = Precision Measuring Equipment Specialist, AGE = Aerospace Ground Equipment, p = persons, f = forms, i:f = items within forms.

Results for the design with tasks nested within methods were similar to those of the crossed design. There was considerable variation across jobs in $\sigma^2_{t:m}$ and $\sigma^2_{i:t:m}$, but little variation in σ^2_{pm}.

These low variance components for the person-by-method interaction indicated that incumbents were not differentially ranked on their performance for the two WTPT methods. This means that incumbents were similarly ranked regardless of method and suggests that interview testing is an acceptable surrogate for hands-on performance testing.



Note. JEM = Jet Engine Mechanic, ACS = Avionic Communications Specialist, ATC = Air Traffic Control Operator, ISRO = Information Systems Radio Operator, ALS = Aircrew Life Support, PS = Personnel Specialist, PMEL = Precision Measuring Equipment Specialist, AGE = Aerospace Ground Equipment.

Figure 4. **Generalizability Coefficients Within Rating Sources for Eight Occupational Specialties.**

D Study Results, WTPT Data

D study analyses were based on the crossed design, because this design permitted assessment of a greater number of effects. D study results for each specialty are displayed graphically in Figure 5.

Unlike the D study results for the rating data, changes in specifications of measurement conditions produced considerable variations in the resulting generalizability curves. Just as increasing the number of rating sources reduced the variance components and improved the generalizability of ratings, using both the hands-on and interview components improved the generalizability of WTPT scores. In general, scores averaged over both methods using a small number of items and a small number of tasks were more generalizable than scores on a single method with a substantially greater number of tasks or items.

Table 3. **Estimated Variance Components for G Study of WTPT Scores with Tasks and Methods Crossed**

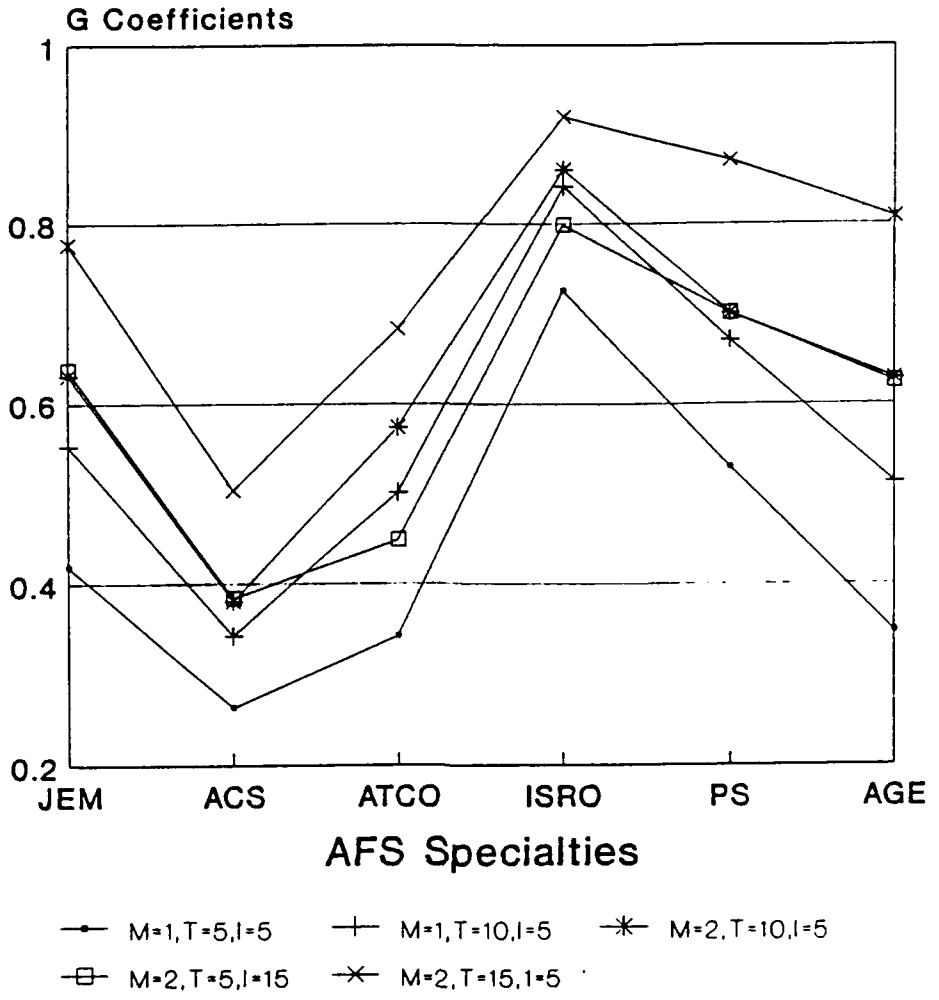
Effect	Job					
	JEM σ^2	ACS σ^2	ATC σ^2	ISRO σ^2	PS σ^2	AGE σ^2
p	.008	.006	.007	.032	.019	.006
m	.013	.014	.000	.000	.003	-.001
t	.000	.016	.008	.007	-.005	.004
i:t	.000	.017	.010	.005	-.003	-.008
mt	.001	.000	.000	.000	.016	.022
pm	.002	.007	.001	.000	-.014	.000
pt	.008	.025	.034	.028	-.014	.001
p(i:t)	.009	.032	.073	.012	.000	-.004
pmt	.012	.008	.007	.020	.078	.020
m(i:t)	.029	.014	.009	.002	.021	.063
pm(i:t)	.127	.074	.065	.052	.094	.149

Note. JEM = Jet Engine Mechanic, ACS = Avionic Communications Specialist, ATC = Air Traffic Control Operator, ISRO = Information Systems Radio Operator, PS = Personnel Specialist, AGE = Aerospace Ground Equipment, p = persons, m = methods, t = tasks, i:t = items within tasks.

Table 4. **Estimated Variance Components for G Study of WTPT Scores with Tasks Nested in Methods**

Effect	Job							
	JEM σ^2	ACS σ^2	ATC σ^2	ISRO σ^2	ALS σ^2	PS σ^2	PMEL σ^2	AGE σ^2
p	.008	.013	.007	.029	.018	.038	.004	.011
m	.013	.001	-.001	-.001	.004	-.007	.004	-.006
t:m	.003	.014	.012	.008	.026	.013	.010	.036
i:t:m	.020	.030	.032	.009	.037	.008	.037	.053
pm	.001	-.001	-.002	-.003	-.001	-.031	-.001	-.006
p(t:m)	.019	.032	.018	.051	.027	.051	.011	.037
p(i:t:m)	.144	.108	.128	.080	.119	.078	.095	.126

Note. JEM = Jet Engine Mechanic, ACS = Avionic Communications Specialist, ATC = Air Traffic Control Operator, ISRO = Information Systems Radio Operator, ALS = Aircrew Life Support, PS = Personnel Specialist, PMEL = Precision Measuring Equipment Specialist, AGE = Aerospace Ground Equipment, p = persons, m = methods, t = tasks, i:t = items within tasks.



Note. JEM = Jet Engine Mechanic, ACS = Avionic Communications Specialist, ATC = Air Traffic Control Operator, ISRO = Information Systems Radio Operator, ALS = Aircrew Life Support, PS = Personnel Specialist, PMEL = Precision Measuring Equipment Specialist, AGE = Aerospace Ground Equipment, M = methods, T = tasks, I = items.

Figure 5. **G Coefficients for WTPT Scores for Six Occupational Specialties.**

Inspection of Figure 5 reveals that the greatest levels of generalizability were obtained for Information Systems Radio Operators, Personnel Specialists, and Aerospace Ground Equipment incumbents. For these specialties, G coefficients above .750 can be obtained with 15 tasks, each with 10 steps, assessed by both hands-on and interview methods. G coefficients were considerably lower in the other specialties. The lowest levels of generalizability occurred for Avionic Communications Specialists. Even with scores averaged over two methods, 15 tasks, and 10 steps, $\underline{\varepsilon P^2}$ equaled only .504. Generalizability levels were somewhat higher for the Air Traffic Control Operators, with $\underline{\varepsilon P^2}$ equal to .683 under the same measurement conditions. It is clear that for these specialties, the WTPT should be constructed with as many items and tasks as feasible. In addition, it is worth noting that generalizability coefficients varied over occupations, making summary conclusions about population estimates of the generalizability of the WTPT suspect. It is safe to assert, however, that the WTPT produces adequate generalizability coefficients regardless of specialty.

Substitutability Design Results

G study estimated variance components, as well as D study estimates of εP^2 for the substitutability design are presented in Table 5. The substitutability design reflects variability in individuals' performance scores across proficiency ratings and WTPT scores for the first four columns of the table, and across ratings, WTPT scores, and job knowledge test scores for the latter three columns. D study estimates are presented for two sets of measurement conditions: a single method of assessing 15 tasks and scores averaged over all three methods, each assessing 15 tasks.

Table 5. **G and D Study Results for Substitutability Design using Supervisor Ratings, WTPT Scores, and Job Knowledge Test Scores**

	Job:						
	JEM	ACS	ATC	ISRO	ALS	PS	AGE
Effect	σ^2	σ^2	σ^2	σ^2	σ^2	σ^2	σ^2
Persons (p)	.016	.012	.007	.031	.617	.087	.063
Methods (m)	3.202	3.196	2.801	4.219	7.767	12.054	6.374
Tasks (t)	-.002	.002	.006	.002	.068	.089	.150
mt	.033	.021	.042	.023	.344	.404	.974
pm	.130	.126	.149	.086	2.245	.327	.323
pt	-.002	.002	.006	.002	.023	.085	.022
pmt	.144	.188	.181	.137	1.900	2.697	1.882
εP^2 when:							
m = 1, t = 15	.104	.076	.044	.244	.207	.178	.126
m = 3, t = 15	.259	.198	.120	.491	.439	.393	.301

Note. JEM = Jet Engine Mechanic, ACS = Avionic Communications Specialist, ATC = Air Traffic Control Operator, ISRO = Information Systems Radio Operator, ALS = Aircrew Life Support, PS = Personnel Specialist, AGE = Aerospace Ground Equipment, p = persons, m = methods, t = tasks.

In no instance were performance scores generalizable over the evaluation methods. In general, only a little over a third of the observed variance in individuals' scores can be attributed to universe score variance (or individual differences). Even in those specialties with the greatest levels of generalizability, these values were well below acceptable levels. A substantial threat to the generalizability coefficients were the large values for the σ^2_{pm} term, indicating that incumbents were differentially ordered by methods.

IV. DISCUSSION

The investigations summarized above supported the JPM Project by applying generalizability theory to three questions: (a) What is the psychometric quality of the performance measures collected as part of the project? (b) How can the scope of the measurement system be reduced without sacrificing the dependability (reliability) of scores? (c) To what extent can conclusions reached on the tested specialties be applied to additional AFSs? In addition, a summary of generalizability theory should address other issues such as the types of research problems for which G theory is and is not suitable, and other research questions generated by G theory results. Each of these issues is discussed in detail below.

Psychometric Quality of Performance Measures

The results of the G and D studies of WTPT scores suggest that WTPTs yield dependable proficiency scores under a variety of conditions. Although the size of the generalizability coefficients varied from specialty to specialty, all coefficients were adequate when at least 10 tasks were assessed. Because the variance component for the persons-by-methods interaction (σ^2_{pm}) was extremely small, it can be concluded that the interview method is an acceptable substitute for the more extensive hands-on method. (Detailed correlational analyses indicating the magnitude of the relationships between the two methods can be found in Hedge, Teachout, & Laue, 1990). Though G coefficients are higher when two methods are used rather than one, the increase in dependability is primarily the result of the added number of tasks. The WTPT is a reliable method of assessing incumbent proficiency.

Conclusions regarding the quality of the proficiency ratings are more problematic. For six of the specialties, generalizability coefficients were greater than .70 when scores were averaged over three sources, at least two forms, and at least eight items. Generalizability coefficients for the other two specialties were only slightly lower. Thus, under such measurement conditions, over two-thirds of the observed variance in scores can be attributed to individual differences. At the same time, reasonable levels of generalizability were obtainable only when multiple forms were used to collect ratings *and* when ratings were averaged over all three sources. Together, these findings confirm the view of ratings as a perceptual phenomenon (Hunter & Hirsch, 1989) in which raters contaminate objective observations of performance with their own perceptual biases and the demands of the particular rating system.

At the same time, D study analyses, within rater level, revealed much higher levels of generalizability with fewer restrictions of measurement conditions. For example, only two 8-item forms were needed for supervisory and peer ratings to produce generalizability coefficients greater than .80 in most specialties. When only a single form was used, nearly all rater sources produced generalizability coefficients greater than .60.

More importantly, the relatively high variance components within sources, coupled with the large σ_{ps}^2 term, suggest that ratings are very dependable within sources but differ considerably (as to how ratees are ranked) across sources. If ratings are to be retained as a criterion for validating the ASVAB, then it will be necessary to collect and combine ratings from all three sources. Otherwise, the question arises as to exactly what is being validated (Wallace, 1965).

Other Measures as Surrogates for WTPT Scores

Evidence of the adequacy of proficiency ratings and job knowledge test scores as surrogates of the WTPT comes from G and D studies of the substitutability design. Regardless of whether scores are averaged across sources or considered separately for each source, there is very little convergence among WTPT scores, ratings, and job knowledge test scores. Thus, task proficiency ratings and job knowledge test scores are not adequate substitutes for the WTPT.

The Need for Research in Additional Specialties

The observed results were very consistent across the eight specialties studied. This applies to both G and D study results of ratings and the substitutability design, and to the attainment of an adequate level of generalizability of the WTPT scores. It is reasonable to conclude that if the same methodologies used to design these measurement systems and to collect data were applied to other specialties, similarly acceptable levels of reliability would be found.

Thus, there would appear to be little need to continue collecting and assessing WTPT or rating data in additional specialties for the sole purpose of estimating their psychometric quality. As seen below, however, there are a number of important questions about criterion measurement which have not been answered by the studies to date. Additional research specifically designed to address these questions is warranted.

Suitable Applications of G Theory

G theory is well suited to the issues addressed immediately above--whether a particular measurement system yields reliable scores, and how that system must be configured to ensure dependability. At the same time, the substitutability analyses were a poor fit of substantive issues and empirical capabilities. Although the G study analyses confirmed that ratings and job knowledge tests were not acceptable substitutes for WTPTs (i.e., the σ_{pm}^2 term was large), neither G study nor D study analyses could determine why incumbents were differentially ranked by methods, or what could be done (from a design perspective) to improve generalizability. A similar limitation occurred in analyses of within-source and between-source differences in performance ratings. The size of the σ_{ps}^2 effect was so large that assumptions of the equivalency of sources (and therefore the appropriateness of the design) are open to question. In either case, the lack of agreement could have been identified by simply inspecting the (low) method inter-correlations. The reasons for the differences require meaningful, theoretically complex studies which may not be easily captured by generalizability analyses.

In future research in performance measurement, G theory may be an appropriate tool for investigating the importance of other measurement parameters. For example, the number of WTPT administrators and their years of experience could be treated as facets in another G study. Ideally, research questions can be framed in G theory terms at the outset of the investigation, rather than applying G theory as a form of analysis after data collection. Restating research questions as G theory problems can help clarify important measurement issues, and ensure that the resulting data can in fact be analyzed using G theory (Cronbach et al., 1972). In addition, G theory will provide a more realistic treatment of measurement error by simultaneously assessing multiple sources of error.

G theory can be applied to Air Force research areas other than Job Performance Measurement. One such area is training evaluation. Two central purposes for evaluating training programs are: (a) validation of the change in either the knowledge states or the performance capacity of trainees; and 2) validation of the training system in which specific decisions were made about the content and process of the particular training course (Kraiger, Salas, & Ford, 1990). Trainee changes in knowledge or performance capacity are ideally assessed through pre- and post-designs and highly reliable criterion measures. Just as G theory can be applied to the assessment of the reliability of performance measures, so too can it be applied to criterion measures for training evaluation. In fact, this was one of the first uses of generalizability theory in educational settings (Cardinet, Tourneur, & Allal 1976; Hopkins, 1983). Some issues that can be addressed with G theory are the number of job knowledge test items necessary for adequate reliability; trainer or instructor effects; training method effects; transfer of training; and decay or maintenance of job skills. For example, one facet for a G study design could be measurement occasions, with the conditions including occasions sampled from different settings (i.e., training vs. on-the-job, or on-the-job at increasingly long periods of time since training). Low estimated variance components for terms containing the interaction of other effects with occasions would indicate that learned skills transfer, or that those skills are maintained on the job.

A second important reason to conduct training evaluation is to provide feedback regarding the outcomes of decisions made during the design of training content or processes. Important questions here concern matters such as whether training outcomes vary as a function of the

instructor, the length of the training program, or the use of instructional media (e.g., programmed instruction vs. videotape presentation). To the extent that the organization has the capability to test these parameters, G theory can be extremely useful for generating research designs which answer these questions.

Conclusions

1. G theory is an appropriate tool for data analysis on the JPM because it offers a versatile and realistic portrayal of measurement error, it forces the researcher to explicitly address important measurement issues, and it permits predictions of the reliability of measurement under a host of possible measurement conditions not currently employed.

2. Proficiency ratings may be adequate criteria for validation purposes. However, if ratings are to be used as criteria for the validation of the ASVAB, then ratings should be collected and averaged over all three sources. Although individual rating sources yield generalizable ratings, these ratings do not agree with ratings from other sources. The construct validity of ratings from a single source would be suspect.

3. The WTPT is an extremely reliable measure of proficiency and should be used for validation purposes. Less extensive versions (e.g., fewer hands-on measures and more interview measures) of the WTPT will yield dependable scores, provided at least 10 to 15 tasks are sampled.

4. Proficiency ratings and job knowledge test scores should not be considered surrogates for the WTPT. They represent different aspects of the total criterion space. Though each methodology is reliable and dependable in and of itself, there is little overlap in the substantive universes assessed by each. Thus, all three measures can be considered "correct," even though they are essentially unrelated.

Recommendations

1. Additional research should be conducted on the construct validity of the performance measures. Some of this work has already been initiated by the Air Force (e.g., Kraiger & Teachout, 1990), but additional work is needed to explain the lack of agreement among different evaluation measures and among different rating sources.

2. Further work should be done to refine the development and administration of the WTPTs. Even though these measures yield dependable scores, it is noted that the residual term (containing random error variance) was still large in many specialties. This suggests that attention to specific aspects of the testing environment (e.g., training of administrators) can yield even more reliable scores.

3. Information about the generalizability of measures should be combined with other information in making decisions about the usefulness of various criterion measures. For example, the cost of each measure can be expressed as a function of the time and expense necessary to develop the measure and/or collect data using the measure. The benefit of a set of measures can be expressed as a function of their generalizability levels and any possible attenuating effects on the relationship of these measures to the ASVAB.

4. G theory should be applied to other research agendas such as training evaluation. Ideally, considerations of appropriate designs for G theory investigations should drive research planning, instead of using G theory simply as a sophisticated tool for data analysis.

REFERENCES

Bentley, B.A., Ringenbach, K.L., & Augustin, J.W. (1989, May). *Development of Army job knowledge tests for three Air Force specialties* (AFHRL-TP-88-11, AD-A208 245). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Borman, W.C. (1974). The ratings of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance, 12*, 105-124.

Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.

Brennan, R.L., & Kane, M.T. (1979). Generalizability theory: A review. In L.J. Fryans, Jr. (Ed.), *Generalizability theory: Inferences and practical applications*. San Francisco, CA: Jossey-Bass.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement, 13*, 119-135.

Crick, J.E., & Brennan, R.L. (1982). *GENOVA: A generalized analysis of variance program (FORTRAN IV computer program and manual)*. Dorchester, MA: Computer Facilities, University of Massachusetts.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley.

Cronbach, L.J., Rajaratnam, N., & Gleser, B. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*, 137-163.

Gillmore, G.M. (1979, March). *An introduction to generalizability theory as a contributor to evaluation research*. Washington University, Seattle, WA: Educational Assessment Center.

Gillmore, G.M. (1983). Generalizability theory: Application to program evaluation. In L.J. Fryans, Jr. (Ed.), *Generalizability theory: Inferences and practical applications*. San Francisco, CA: Jossey-Bass.

Hedge, J.W., & Teachout, M.S. (1986, November). *Job performance measurement: A systematic program of research and development* (AFHRL-TP-86-37, AD-A174 175). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Hedge, J.W., Teachout, M.S., & Laue, F.J. (1990, November). *Interview testing as a work sample measure of job proficiency* (AFHRL-TP-90-61, AD-A228 054). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Hopkins, K.D. (1983). Estimating reliability and generalizability coefficients in two-facet designs. *Journal of Special Education, 17*, 371-375.

Hunter, J.E., & Hirsch, H.R. (1989). Applications of meta-analysis. In C.L. Cooper & I.T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology*. New York, NY: Wiley.

Klimoski, R.J., & London, M. (1974). Role of the rater in performance appraisal. *Journal of Applied Psychology, 59*, 445-451.

Kraiger, K. (1989, April). *Generalizability theory: An assessment of its relevance to the Air Force job performance measurement project* (AFHRL-TP-87-70, AD-A207-107). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Kraiger, K. (1990a, February). *Generalizability of performance measures across four Air Force specialties* (AFHRL-TP-89-60, AD-A220 821). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Kraiger, K. (1990b, March). *Generalizability of walk-through performance tests, job proficiency ratings and job knowledge tests across eight Air Force specialties* (AFHRL-TP-90-14, AD-A225 011). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Kraiger, K., Salas, E., & Ford, J.K. (1990). *Comprehensive training evaluation*. Unpublished manuscript, Denver, CO: University of Colorado at Denver.

Kraiger, K., & Teachout, M.S. (1990). Generalizability theory as construct-related evidence of the validity of job performance ratings. *Human Performance, 3*, 19-35.

Littlefield, J.H., Murrey, A.J., & Garman, R.E. (1977, April). *Assessing the generalizability of clinical rating scales*. Paper presented at the meeting of the American Educational Research Association.

McHenry, J.J., Hoffman, R.G., & White, L.A. (1987, April). A generalizability analysis of peer and supervisory ratings. In G. Laabs (Chair), *Applications of generalizability theory to military performance measurement*. Symposium at the annual meeting of the American Educational Research Association, Washington, DC.

Searle, S.R. (1971). *Linear models*. New York, NY: Wiley.

Shavelson, R.J. (1986, July). *Generalizability of military performance measurements: I. Individual performance*. Paper prepared for the Committee on the Performance of Military Personnel and the Commission on Behavioral and Social Sciences and Education, National Research Council, and National Academy of Sciences.

Shavelson, R.J., & Webb, N.M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133-161.

Wallace, S.R. (1965). Criteria for what? *American Psychologist, 20*, 411-417.

Webb, N.M., Shavelson, R.J., Kim, K.S., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinist mates. *Military Psychology, 1*, 91-110.

Webb, N.M., Shavelson, R.J., Shea, J., & Morello, E. (1981). Generalizability of General Educational Development ratings of jobs in the U.S. *Journal of Applied Psychology*, *66*, 186-191.

Wigdor, A.K., & Green, B.F., Jr. (1986). *Assessing the performance of enlisted personnel: Evaluation of a joint-service research project*. Washington, DC: National Academy Press.

• U. S. GOVERNMENT PRINTING OFFICE: 1991--761-052/40060